Prediction of Crop Yield

DEPARTMENT OF CHEMICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI

November 28, 2024

Shobhit Kumar 2023CH70769

Deepanshu 2023CH10881

Nitin Yadav 2023CH11150

Rishabh Dewangan 2023CH70813

Abstract

Yield prediction can be really helpful for farmers to predict their profit. We used Multivariable Linear regression to stablish relationship between temperature, rainfall and yield with other environmental factors. Found outstanding correlation coefficient which clearly shows yield mainly depends on temperature, rainfall.Research can be extended using more variables like food price index(FPI), minimum support price(MSP).

1 Introduction

One of the most important occupations that serve human beings has always been agriculture,both in terms of livelihood and employment. The food source of the poor is growing bad due to the significant increase in the population, which need to be improved. Therefore, there has been great requirement of adoption of new and better practices of crops. The rapid growth in population has a significant impact on the environment, with noticeable environmental damage.

Historical advancements in agriculture clearly demonstrate human's capability in meeting food demands, even in the phase of increasing population growth. Between human beings and their environment a balanced relationship should be made in order to lead a sustainable life.

An examination of the contribution of agriculture to the national income and its share in export for a period of 50 years makes clear that the share of agriculture in the national income and in the total export is declining consistently.

Yield prediction is one of the most critical issues faced in the agricultural sector. Farmer's who have lack of knowledge, uncertainties in the weather conditions and seasonal rainfall policies, depletion of nutrition level of soils, fertilizer availability and cost, pest control, post–harvest loss and other factors leads to decrease in the production of the crops

Analysis of data for the research purpose on decision making and problem solving requires regression analysis. This requires analysis of multiple variables or objects for efficient prediction of yield. We have considerd multiple factors like region(North,South,East or West), Soil Type (Sandy,Clay,Silt,Loam,Peaty or Chalky), Weather Conditions like Rainy, Sunny or Cloudy, Fertilzer(used or not), Irrigation(Used or not), temperature, rainfall, days to harvest and many more.

2 Agriculture in india

The agricultural history of India has been documented since 1100 BC during the Rig Veda period.[ref: https://en.wikipedia.org/wiki/ Agriculture in India].

Studies show that agricultural production of India have increased by improving the farm productivity and grain storing infrastructure , just not only to feed growing population but also export them globally.

In 2011 , during monsoon season , Indian agricultural have increased by 6.4 per from previous year , hitting all time record of 85.9 million tons of wheat production.

In the same time , production of rice have increased by 7 per , touching a all time record of 95.3 million tons production.

In 2013, India exported agricultural products worth 39 billion dollars, making it the seventh-largest agricultural exporter in the world and the sixth-largest net exporter. It was also ranked second globally in farm output, reflecting its strong agricultural productivity. During this time, agriculture, forestry, and fisheries contributed 13.7 per to India's total GDP, highlighting the sector's importance to the economy, particularly in rural areas. These techniques were applied to predict crop yields, analyze soil health, forecast market trends, detect plant diseases, and improve irrigation management.

India's agricultural contribution to the total GDP is decreasing due to rapid economic growth in India. Agricultural sector plays a significant role in the socio-economic growth of India and is still the biggest economic sector. India exported 39 billion dollars worth of agricultural products in 2013 and making it the seventh largest agricultural exporter worldwide and the sixth largest net exporter.

India plays a significant role in global agriculture, being one of the largest suppliers of essential crops like rice, wheat, cotton, and sugar. The country has exported over 2 million metric tons of wheat and 2.1 million metric tons of rice, catering primarily to markets in Asia and Africa. Additionally, India ranks as the second or third-largest producer globally in a variety of agricultural and related products.

3 Numerical method used

3.1 Gauss Elimination

This method deals with simultaneous linear algebraic equations that can be represented generally as

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \tag{3.1.1}$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \tag{3.1.2}$$

$$\vdots$$

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \tag{3.1.n}$$

As was the case with the solution of two equations, the technique for n equations consists of two phases: elimination of unknowns and solution through back substitution.

3.1.1 Naive Gauss Elimination

The elimination of unknowns was used to solve a pair of simultaneous equations. The procedure consisted of two steps:

1. The equations were manipulated to eliminate one of the unknowns from the equations. The result of this elimination step was that we had one equation with one unknown.

2. Consequently, this equation could be solved directly and the result back-substituted into one of the original equations to solve for the remaining unknown.

This basic approach can be extended to large sets of equations by developing a systematic scheme or algorithm to eliminate unknowns and to back-substitute. Gauss elimination is the most basic of these schemes. This section includes the systematic techniques for forward elimination and back substitution that comprise Gauss elimination. Although these techniques are ideally suited for implementation on computers, some modifi cations will be required to obtain a reliable algorithm. In particular, the computer program must avoid division by zero. The following method is called "naive" Gauss elimination because it does not avoid this problem. Subsequent sections will deal with the additional features required for an effective computer program. The approach is designed to solve a general set of n equations:

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \tag{3.1.1.1a}$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \tag{3.1.1.1b}$$
$$\vdots$$
$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \tag{3.1.1.1c}$$

Forward Elimination of Unknowns. The first phase is designed to reduce the set of equations to an upper triangular system. The initial step will be to eliminate the first unknown, x1, from the second through the nth equations. To do this, multiply Eq. (3.1.11) by a21/a11 to give

$$a_{21}x_1 + \frac{a_{21}}{a_{11}}a_{12}x_2 + \cdots + \frac{a_{21}}{a_{11}}a_{1n}x_n = \frac{a_{21}}{a_{11}}b_1 \quad (3.1.1.2)$$

Now, this equation can be subtracted from Eq. (3.1.1.1b) to give

$$\left(a_{22} - \frac{a_{21}}{a_{11}}a_{12}\right)x_2 + \cdots + \left(a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}\right)x_n = b_2 - \frac{a_{21}}{a_{11}}b_1$$

or

$$a'_{22}x_2 + \cdots + a'_{2n}x_n = b'_2$$

where the prime indicates that the elements have been changed from their original values. The procedure is then repeated for the remaining equations. For instance, Eq. (3.1.1.1a) can be multiplied by a31/11 and the result subtracted from the third equation. Repeating the procedure for the remaining equations results in the following modified system:

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1, & (3.1.1.3a)\\
a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2, & (3.1.1.3b)\\
a'_{32}x_2 + a'_{33}x_3 + \cdots + a'_{3n}x_n &= b'_3, & (3.1.1.3c)\\
&\vdots\\
a'_{n2}x_2 + a'_{n3}x_3 + \cdots + a'_{nn}x_n &= b'_n. & (3.1.1.3d)
\end{aligned}$$

For the foregoing steps, Eq. (3.1.1.2a) is called the pivot equation and a11 is called the pivot coefficient or element. Note that the process of multiplying the first row by a21/a11 is equivalent to dividing it by a11 and multiplying it by a21. Sometimes the division operation is referred to as normalization. We make this distinction because a zero pivot element can interfere with normalization by causing a division by zero. We will return to this important issue after we complete our description of naive Gauss elimination. Now repeat the above to eliminate the second unknown from Eq. (3.1.1.2c) through (3.1.1.1d). To do this multiply Eq. (3.1.1.2b) by a'_{32}/a'_{22} and subtract the result from Eq. (3.1.1.1c). Perform a similar elimination for the remaining equations to yield

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1\\
a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2\\
a''_{33}x_3 + \cdots + a''_{3n}x_n &= b''_2\\
&.\\
&.\\
&.\\
&.\\
a''_{n3}x_3 + \cdots + a''_{nn}x_n &= b''_n
\end{aligned}$$

where the double prime indicates that the elements have been modified twice. The procedure can be continued using the remaining pivot equations. The final manipulation in the sequence is to use the (n - 1)th equation to eliminate the x_{n-1} term from the nth equation. At this point, the system will have been transformed to an upper triangular system

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n = b_1, \quad (3.1.1.4a)$$
$$a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n = b'_2, \quad (3.1.1.4b)$$
$$a''_{33}x_3 + \cdots + a''_{3n}x_n = b''_3, \quad (3.1.1.4c)$$
$$..$$
$$..$$
$$..$$
$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \quad (3.1.1.4d)$$

Back Substitution. Equation (3.1.1.4d) can now be solved for x_n:

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

This result can be back-substituted into the (n - l)th equation to solve for x_{n-1}. The procedure, which is repeated to evaluate the remaining x's, can be represented by the following formula:

$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)} x_j}{a_{ii}^{(i-1)}} \quad \text{for } i = n-1, n-2, \ldots, 1$$

3.2 Regression analysis

Regression analysis is used to analyze and determine the relationship between the response variable (crop yield) and explanatory variables (such as soil type, rainfall, temperature, etc.). Many ecological factors influence crop yield. In this study, we analyze the relationship between crop yield and various factors such as area of cultivation, rainfall, temperature, etc.

3.2.1 Multiple Linear Regression

A useful extension of linear regression is when the outcome (y) depends on two or more independent variables. For example, y can be a combination of x1 and x2, written as:

$$y = a_0 + a_1x_1 + a_2x_2 + e$$

This equation is helpful when fitting experimental data, where the variable being studied is influenced by two other factors. In this case, instead of a straight line, the regression becomes a "plane" in a two-dimensional space.

In previous case , the best values of the coefficients are determined by setting up the sum of the e squares of the residuals,

$$Sr = \sum_{i=1}^{n} (y_i - a_0 - a_1x_{1i} - a_2x_{2i})^2$$

and differentiating with respect to each of the unknown coefficients,

$$\frac{\partial S_r}{\partial a_0} = -2\sum (y_i - a_0 - a_1x_{1i} - a_2x_{2i})$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum (x_{1i}(y_i - a_0 - a_1x_{1i} - a_2x_{2i}))$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum (x_{2i}(y_i - a_0 - a_1x_{1i} - a_2x_{2i}))$$

The coeffi cients yielding the minimum sum of the squares of the residuals are obtained by setting the partial derivatives equal to zero and e.3333

xpressing the result in matrix form as

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{bmatrix}$$

The foregoing two-dimensional case can be easily extended to m dimensions, as in

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m + e$$

where the standard error is formulated as

$$s_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}}$$

and the coefficient of determination is computed as

$$r^2 = \frac{S_t - S_r}{S_t}$$

In the preceding pages, we have introduced three types of regression: simple linear, polynomial, and multiple linear regression. In fact, all three belong to the following general linear least-squares model:

$$y = a_0z_0 + a_1z_1 + a_2z_2 + \cdots + a_mz_m + e$$

where $z_0, z_1, \ldots, z_m$ are $m+1$ basis functions. It can easily be seen how simple and multiple linear regression fall within this model—that is, $z_0 = 1, z_1 = x_1, z_2 = x_2, \ldots, z_m = x_m$. Further, polynomial regression is also included if the basis functions are simple monomials as in $z_0 = x^0 = 1, z_1 = x, z_2 = x^2, \ldots, z_m = x^m$.

Note that the terminology "linear" refers only to the model's dependence on its parameters—that is, the a's. As in the case of polynomial regression, the functions themselves can be highly nonlinear. For example, the z's can be sinusoids, as in

$$y = a_0 + a_1\cos(\omega t) + a_2\sin(\omega t)$$

Such a format is the basis of Fourier analysis described in Chap. 19.

On the other hand, a simple-looking model like

$$f(x) = a_0\left(1 - e^{-a_1x}\right)$$

is truly nonlinear because it cannot be manipulated into the format of Eq. (17.23). We will turn to such models at the end of this chapter.

For the time being, Eq. (17.23) can be expressed in matrix notation as

$$\{Y\} = [Z]\{A\} + \{E\}$$

where $[Z]$ is a matrix of the calculated values of the basis functions at the measured values of the independent variables,

$$[Z] = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix},$$

m is the number of variables in the model, and n is the number of data points. Because $n \geq m+1$, you should recognize that most of the time, $[Z]$ is not a square matrix.

The column vector $\{Y\}$ contains the observed values of the dependent variable

$$\{Y\}^T = [y_1\, y_2\, \cdots\, y_n],$$

the column vector $\{A\}$ contains the unknown coefficients

$$\{A\}^T = [a_0\, a_1\, \cdots\, a_m],$$

and the column vector $\{E\}$ contains the residuals

$$\{E\}^T = [e_1\, e_2\, \cdots\, e_n].$$

As was done throughout this chapter, the sum of the squares of the residuals for this model can be defined as

$$S_r = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} a_j z_{ji} \right)^2 .$$

This quantity can be minimized by taking its partial derivative with respect to each of the coefficients and setting the resulting equation equal to zero. The outcome of this process is the normal equations that can be expressed concisely in matrix form as

$$[[Z]^T[Z]]\{A\} = [[Z]^T\{Y\}].$$

It can be shown that Eq. (17.25) is, in fact, equivalent to the normal equations developed previously for simple linear, polynomial, and multiple linear regression.

Our primary motivation for the foregoing has been to illustrate the unity among the three approaches and to show how they can all be expressed simply in the same matrix notation. The matrix notation will also have relevance when we turn to nonlinear regression in the last section of this chapter.

From Eq. (PT3.6), recall that the matrix inverse can be employed to solve Eq. (17.25), as in

$$\{A\} = [[Z]^T[Z]]^{-1}[[Z]^T\{Y\}].$$

As we have learned in Part Three, this is an inefficient approach for solving a set of simultaneous equations. However, as a simplification strategy, there are a number of contexts where it proves useful to develop and understand the results by employing this approach.

4 Numerical Approach

4.1 Finding data

We downloaded a CSV file from Kaggle for regression analysis and further processing. Below is a sample of the agricultural data used in the study. We filtered our data using excel and observed only for rice further.

Region	Soil	Crop	Rain(mm)	T(°C)	Fer	Irr	Weather	Harvest	Y(t/ha)
West	Sandy	Cotton	897.0772	27.6770	0	1	Cloudy	122	6.5558
South	Clay	Rice	992.6733	18.0261	1	1	Rainy	140	8.5273
North	Loam	Barley	147.9980	29.7940	0	0	Sunny	106	1.1274
North	Sandy	Soybean	986.8663	16.6442	0	1	Rainy	146	6.5176
South	Silt	Wheat	730.3792	31.6207	1	1	Cloudy	110	7.2483
South	Silt	Soybean	797.4712	37.7050	0	1	Rainy	74	5.8984
West	Clay	Wheat	357.9024	31.5934	0	0	Rainy	90	2.6524
South	Sandy	Rice	441.1312	30.8871	1	1	Sunny	61	5.8295
North	Silt	Wheat	181.5879	26.7527	1	0	Sunny	127	2.9437
West	Sandy	Wheat	395.0490	17.6462	0	1	Rainy	140	3.7073

Table 1: Agricultural Data with Weather and Yield Information

4.2 One-Hot Encoding and Normalization

4.2.1 One-Hot Encoding

One Hot Encoding is a method for converting categorical variables into a binary format. It creates new binary columns (0s and 1s) for each category in the original variable. Each category in the original column is represented as a separate column, where a value of 1 indicates the presence of that category, and 0 indicates its absence.

4.2.2 Data Normalization

Converted Rainfall(mm) and Temperature(C) into small values between 0 and 1 to ensure uniformity across all columns. Following is formula for normalization

$$X_i = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Following is Data after encoding and normalization.

Sandy	Clay	Loam	Silt	Peaty	Chalky	Rain_Nor	Temp_Nor	Fert	Irr
0	1	0	0	0	0	0.9919	0.1210	1	1
1	0	0	0	0	0	0.3790	0.6355	1	1
0	0	0	0	0	1	0.8605	0.4903	1	0
0	1	0	0	0	0	0.9287	0.2355	1	1
0	0	0	1	0	0	0.4185	0.5736	1	1
0	0	1	0	0	0	0.0897	0.1285	0	0
0	0	0	0	1	0	0.7841	0.5680	1	0
0	0	0	0	1	0	0.5935	0.9381	0	0
0	1	0	0	0	0	0.7799	0.5505	0	1
0	0	0	0	1	0	0.3095	0.9257	1	0

Rainy	Sunny	Cloudy	Harvest	Reg_N	Reg_E	Reg_W	Reg_S	Yield
1	0	0	140	0	0	0	1	8.53
0	1	0	61	0	0	0	1	5.83
0	1	0	115	0	1	0	0	5.84
1	0	0	71	0	0	1	0	7.76
0	1	0	67	0	0	0	1	6.41
0	0	1	83	1	0	0	0	0.80
1	0	0	145	0	0	1	0	6.34
0	1	0	127	0	1	0	0	4.17
0	1	0	124	0	0	1	0	5.35
0	0	1	64	0	0	0	1	4.66

Table 2: Modified agricultural data

4.3 Matrix X and Y

We converted above Table 2: data into seperate columns and converted each column into csv then using csv file we took input using C++ into following matrix X and Y.Here $X_{1,1}, X_{1,2}, \ldots$ are the first observations of all 18 variables and so on and $Y_1, Y2, ..$ are observations of Yield in tons/hectare. a_0 is constant and $a_1, a2, ..$ are coefficients of all variables.

$$X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,18} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,18} \\ 1 & X_{3,1} & X_{3,2} & \cdots & X_{3,18} \\ 1 & X_{4,1} & X_{4,2} & \cdots & X_{4,18} \\ 1 & X_{5,1} & X_{5,2} & \cdots & X_{5,18} \\ 1 & X_{6,1} & X_{6,2} & \cdots & X_{6,18} \\ 1 & X_{7,1} & X_{7,2} & \cdots & X_{7,18} \\ 1 & X_{8,1} & X_{8,2} & \cdots & X_{8,18} \\ . & . & . & \cdots & . \end{bmatrix}, A = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ . \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ . \end{bmatrix}$$

4.4 Solving For Coefficients

Now A matrix can be solved using regression formula

$$A = (X^T X)^{-1}(X^T Y)$$

To solve for A we used gauss elimination

$$PA = Q$$

$$here, P = X^T X, Q = X^T Y$$

4.5 Statistical Evaluation

Calculated S_r using formula

$$S_r = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Calculated S_t using formula

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Calculated r^2 using formula

$$r^2 = 1 - \frac{S_r}{S_t}$$

Calculated adjusted r^2 using formula

$$r^2_{\text{adj}} = 1 - \left(\frac{(1 - r^2)(n - 1)}{n - k - 1} \right)$$

here n is number of observations and k is number of independent variables.

5 Results and Discussions

We manipulated raw data using excel, developed Code for Regression, Statistical Evaluation using C++. Plotted different graphs and analysed them and matched them with our results obtained using code.

5.1 Crop Yield Trends

Following observations we found by analysing and plotting data.

5.1.1 Impact of Irrigation and Fertilizers

Crop Yield when Fertilizers was used but Irrigation was not.

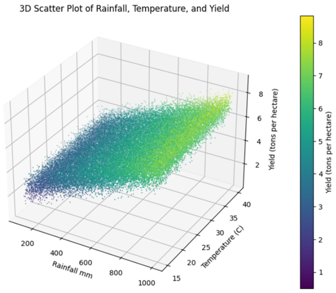

Figure 1: Rainfall, Temperature and Yield with Fertilizer without Irrigation.

Crop Yield when Fertilizers was not used but Irrigation was.

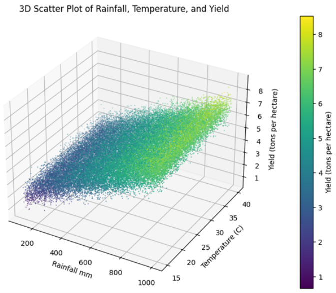

Figure 2: Rainfall, Temperature and Yield without Fertilizer with Irrigation.

Crop Yield when both Fertilizers, Irrigation was not used.

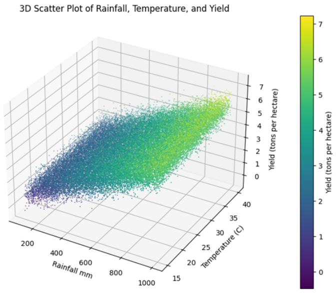

Figure 3: Rainfall, Temperature and Yield without Fertilizer and Irrigation.

Crop Yield when both Fertilizers and Irrigation are used.

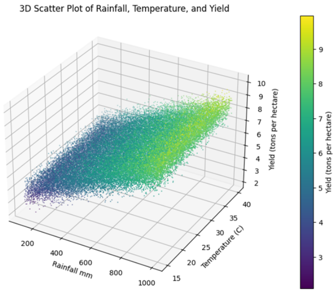

Figure 4: Rainfall, Temperature and Yield with Fertilizer and Irrigation.

All 4 cases in single plot. Showing that overall Yield is highest when both irrigation and fertilizer are used, yield of only fertilzer is greater than yield of only irrigation and at last yield will be lowest when neither irrigation nor fertilzer was used.

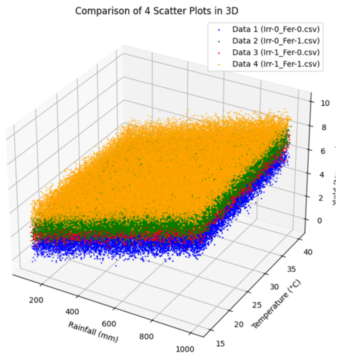

Figure 5: Rainfall, Temperature and Yield.

5.1.2 Impact of Region

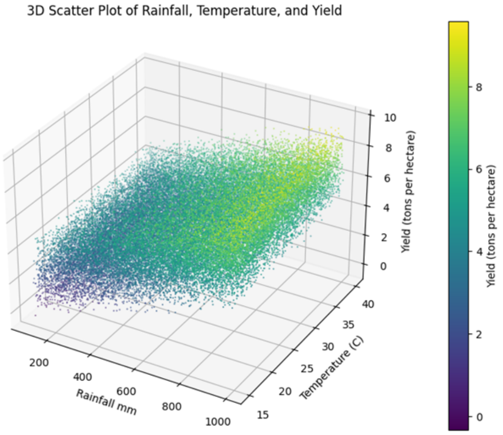

Figure 6: Region North.

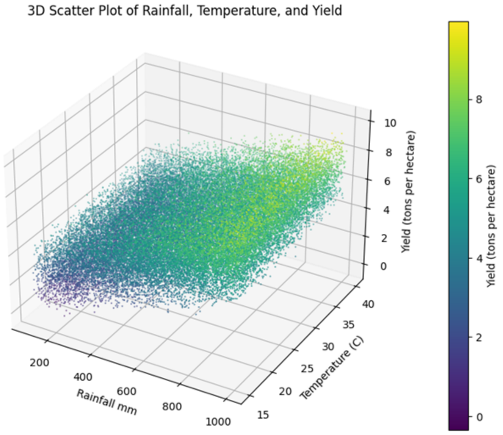

Figure 7: Reg South.

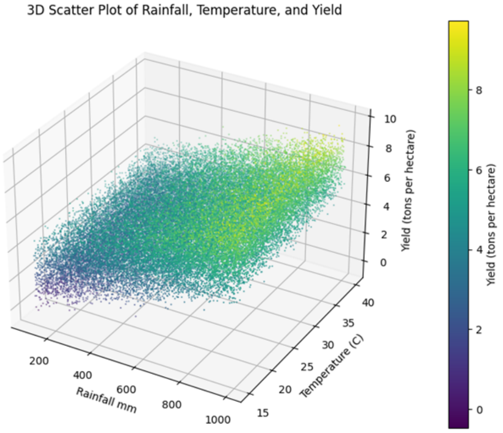

Figure 8: Region East.

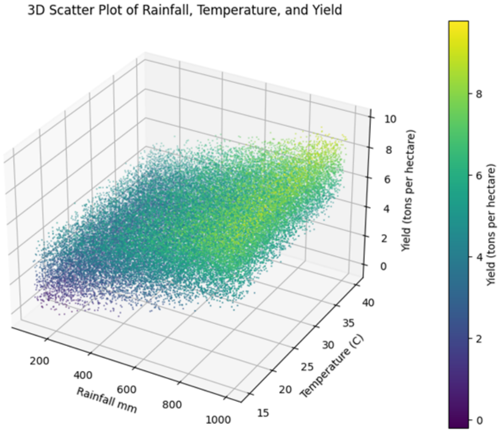

Figure 9: Reg West.

We found that Yield Doesn't depend much on region from above figure it is easy to say. Also in our regression we found Coefficients of each region North, South, East, West equals to each other. Also a common plot shows that all 4 regions have same plane given below.

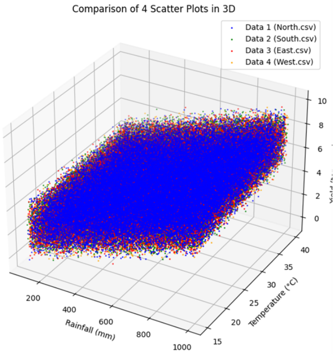

Figure 10: Rainfall, Temperature and Yield.

5.1.3 Impact of days to harvest

The number of days taken for the crop to be harvested after planting. After observing graph we found that both graph have same plane which means days to harvest don't effect yield of crop also which is verified by our regression in which coefficient of this variable was near about zero.

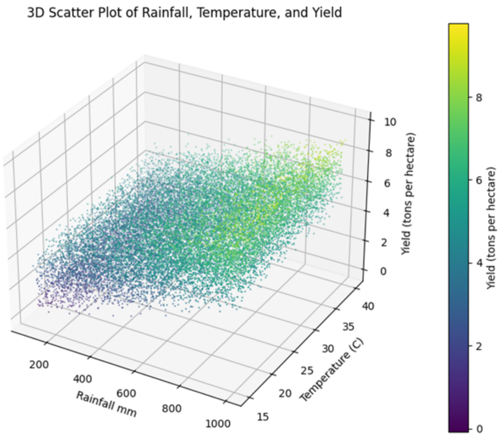

Figure 11: 60-70 days to harvest.

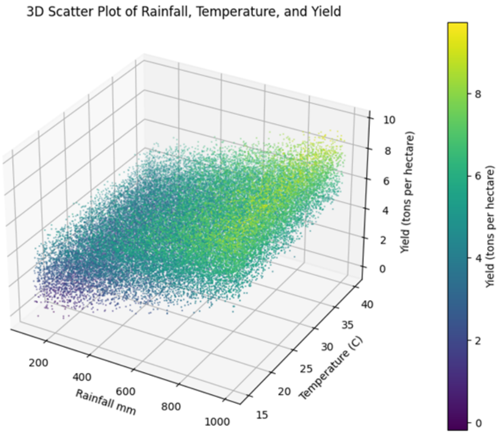

Figure 12: 120-140 days to harvest.

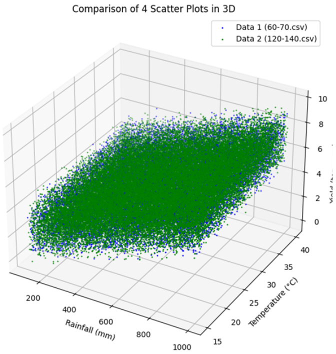

Figure 13: Rainfall, Temperature and Yield for different time to harvest.

5.2 Temperature and Rainfall

We observed in all of the above plots that rate of change of yield wrt Rainfall at constant temperature is more than rate of change of yield wrt temperature at constant rainfall, same result we obtained in our regression analysis in which we found coefficient of rainfall was greater than coefficient of temperature. Since plot is looking like a plane thats why multilinear regression is best choice on given data. Our observations gives us outstanding results with r^2 value equals to 0.91.

5.3 Results of Statistical Evaluations

We performed linear regression and found coefficients of all independent variable

5.3.1 Regions

Region_East (X_1): $a_1 = -26.12$ Being in the East decreases the yield by approx. 26 tons/hectare.
Region_West (X_2): $a_2 = -26.12$ Being in the West decreases the yield by approx. 26 tons/hectare.
Region_North (X_3): $a_3 = -26.12$ Being in the North decreases the yield by approx. 26 tons/hectare.
Region_South (X_4): $a_4 = -26.12$ Being in the South decreases the yield by approx. 26 tons/hectare.
All 4 regions have approx. same coefficients which means region doesn't effect Yield much.

5.3.2 Soil Types

Soil_Type_Chalky (X_5): $a_5 = -18.54$ Chalky soil decreases the yield by approximately 18.54 tons/hectare.
Soil_Type_Clay (X_6): $a_6 = -18.53$ Clay soil decreases the yield by approximately 18.53 tons/hectare.
Soil_Type_Loam (X_7): $a_7 = -18.53$ Loam soil decreases the yield by approximately 18.53 tons/hectare.
Soil_Type_Peaty (X_8): $a_8 = -18.54$ Peaty soil decreases the yield by approximately 18.54 tons/hectare.
Soil_Type_Sandy (X_9): $a_9 = -18.54$ Sandy soil decreases the yield by approximately 18.54 tons/hectare.
Soil_Type_Silt (X_{10}): $a_{10} = -18.53$ Silt soil decreases the yield by approximately 18.53 tons/hectare.
All 6 soil types have approximately same coefficient which means soil type doesn't effect yield much. But if we mix some soils there can be better yield if there is no other factors.

5.3.3 Days to Harvest

Days_to_Harvest (X_{11}): $a_{11} = 0.0001$ Each additional day to harvest slightly increases the yield by 0.0001 tons/hectare.

5.3.4 Fertilizer and Irrigation

Fertilizer_Used (X_{12}): $a_{12} = 1.50$ Using more fertilizer increases the yield by approximately 1.50 tons/hectare per unit. Irrigation_Used (X_{13}): $a_{13} = 1.20$ Using more irrigation increases the yield by approximately 1.20 tons/hectare per unit.

5.3.5 Environmental Factors

Rainfall_mm_Normalized (X_{14}): $a_{14} = 4.50$ Increased normalized rainfall increases the yield by approximately 4.50 tons/hectare per unit. Temperature_Celsius_Normalized (X_{15}): $a_{15} = 0.49$ Higher normalized temperature increases the yield by approximately 0.49 tons/hectare per unit. Rainfall effects more on Yield as compared to temperature.

5.3.6 Weather Conditions

Weather_Condition_Cloudy (X_{16}): $a_{16} = 83.00$ Cloudy weather increases the yield by approximately 83.00 tons/hectare. Weather_Condition_Rainy (X_{17}): $a_{17} = 83.00$ Rainy weather increases the yield by approximately 83.00 tons/hectare. Weather_Condition_Sunny (X_{18}): $a_{18} = 82.99$ Sunny weather increases the yield by approximately 82.99 tons/hectare. Each weather have same coeff. that's why there is no extra impact of any weather condition.

5.3.7 Baseline

Intercept (a_0): $a_0 = -37.55$ The baseline yield, when all variables are zero, is -37.55 tons/hectare.

5.3.8 Correlation Coefficient

Correlation coefficient shows relationship between predicted values and actual values of Yield. We calculated correlation coefficient which found to be equals to 0.91 and Adjusted r^2 equals to 0.91. Fertilizer coeff. is greater than Irrigation coeff.. Yield is independent of region, soil, weather condition. Days to harvest also has almost zero impact on yield.

6 Path forward

Our study will further investigate the integration of ensemble techniques, such as combining regression models with decision trees and other machine learning algorithms, to enhance the accuracy and reliability of predictions across diverse agricultural conditions. Efforts will also aim to expand the dataset by incorporating temporal trends, including multi-seasonal data, to improve the model's adaptability for long-term applications.

7 Conclusion

We used linear regression algorithm along with gauss elimination to predict the relationship between region, soil-type, days to harvest, weather condition, temperature, rainfall, fertilizer used, irrigation used, and we found results with correlation coefficient equal to 0.91. We found region, weather condition, days to harvest does not impact much on Yield of Rice crop. Only using fertilizer can increase yield much more than just only using irrigation. Using both fertilizer and irrigation can increase yield much significantly. Higher rainfall and temperature also can increase yield but they are not in our control but they have significant impact on yield.

8 Self-Assessment

Level 0 We did complete literature assessment of the formulae, regression model, numerical methods.
Level 1 We developed working code in C++ built linear regression algorithm, gauss elimination algorithm. Predicted coefficient of all variables. Found linear relationship between all variables and yield. Developed code for statistical evaluations.
Level 2 Our regression model not just show impacting variables but also showing factors which don't effect yield of rice crop which allow us to focus more on the depending factors. Our correlation coefficient found to be equals to 0.91 which is more than our reference journal article whose correlation coefficient value was 0.7. We calculated coefficients of each factors and shown how much dependency on each factor. Showing importance of fertilizers and irrigation and which one should be used if we only have one choice to use. We have shown regions, soil-type, weather conditions, days to harvest does not impact yield. Temperature and rainfall are major factors affecting yield. All these results show that our regression model and analysis are of level 2.

9 References

[1]Samuel Oti Attakorah. Agriculture Crop Yield Dataset https://www.kaggle.com/datasets/samuelotiattakorah/agricultu crop-yield/data.
[2]Prof.Jayati Sarkar notes numerical methods in Chemical Engineering. Course Code CLL-113.
[3]V. Sellam* and E. Poovammal. Prediction of Crop Yield using Regression Analysis.Indian Journal of Science and Technology, Vol 9(38), 10.17485/ijst/2016/v9i38/91714, October 2016.
[4]. Kalpana, Shanthi, Arumugam. A survey on data mining techniques in agriculture. International Journal of Advances in Computer Science and Technology. 2014 Aug; 3(8). 2320–2602. numerical methods sc chapra.
[5]. Mucherino A, Papajorgji P, Pardalos PM. A survey of data mining techniques applied to agriculture. Springer–Verlag; 2009 Jun.
[6] Kokilavani S, Geethalakshmi V. Identification of efficient cropping zone for rice, maize and groundnut in Tamil Nadu. Indian Journal of Science and Technology. 2013 Oct; 6(10).

[7]. Chinchuluun A, Xanthopoulos P. Data mining techniques in agricultural and environmental sciences. 26 International Journal of Agricultural and Environmental Information Systems; 2010 Jan-Jun; 1(1):26–40.

[8]. Suraparaju V, Misra B, Singh CD. Machine learning approach for forecasting crop yield based on climatic parameters. International Conference on Computer Communication and Informatics (ICCCI–2014); Coimbatore, India. 2014 Jan 3–5. Figure 1. Yield prediction from AUC for rice Figure 2. Yield prediction from AR for rice. Figure 3.: Yield prediction from FPI for rice Figure 4. Relationship between MSP to FPI. V. Sellam and E. Poovammal Indian Journal of Science and Technology5Vol 9 (38) — October 2016 — www.indjst.org

[9]. Mankar AB, Burange MS. Data mining - An evolutionary view of agriculture. International Journal of Application or Innovation in Engineering and Management.2014 Mar; 3(3):102–5.

[10]. Ramesh D, Vishnu Vardhan B. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering.2013 Sep; 2(9):3477–80.

[11]. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate data analysis. 6th ed. Pearson Education Inc; 2006.

[12]. Patel H, Patel D. A brief survey of data mining techniques applied to agricultural data. International Journal of Computer Applications. 2014 Jun; 95(9):6–8.

[13]. Kumar DA, Kannathasan N. A survey on data mining and pattern recognition techniques for soil data mining. IJCSI International Journal of Computer Science Issues. 2011 May; 8(3):422–8.

[14]. Kalpana, Shanthi, Arumugam. A survey on data mining techniques in agriculture. International Journal of Advances in Computer Science and Technology. 2014 Aug; 3(8). 2320–2602. numerical methods sc chapra.